图书情报工作
第65卷 第22期 2021年11月

基于词和实体标注的古籍数字人文知识库的构建与应用*
——以《资治通鉴·周秦汉纪》为例

■ 常博林[1] 万晨[2] 李斌[1] 陈欣雨[1] 冯敏萱[1] 王东波[3]
[1] 南京师范大学文学院 南京 210097 [2] 复旦大学中国语言文学系 上海 200433
[3] 南京农业大学信息管理学院 南京 210095

摘 要:[目的/意义]探索能够实现基于词和实体的检索与知识挖掘的人文知识库构建方法。[方法/过程]以《资治通鉴·周秦汉纪》为例,对68卷60万字的文本自动分词与词性标注之后,人工标注文本中的人物、地点GIS、时间等实体信息,实现基于词和实体的全文检索和地图检索系统;利用同现信息,统计出人物关系与人物游历信息;进而使用TF-IDF方法,通过时间序列分析,挖掘出多事之秋、风云人物、风云之地等结果。[结果/结论]基于词和实体的深度信息标注,能够解决缺乏词界、同名异指和异名同指的检索难题,更可以为古籍多角度的知识发掘与知识服务提供基础支撑。

关键词:《资治通鉴》 数字人文 知识挖掘 古籍检索 古文信息处理
分类号:G250
DOI:10.13266/j.issn.0252-3116.2021.22.014

1 引言

中国的古籍文献数量庞大、包罗万象,是研究中国的语言、文学、历史、文化等方面的宝藏。上世纪末以来,古籍的电子化和基于字的全文检索系统已经取得了许多重大进展,形成了一大批可用的电子数据库[1]。而随着数字人文技术的兴起[2],国际史学界已经开始从纸质文献的文字历史叙述转变为结构化历史数据库。希罗多德的历史数据库[3]和中国历史人物传记数据库(CBDB)[4]都尝试把历史的要素,诸如时间、人物、地点进行详细的描述与关联,形成可检索、可视化的历史数据平台。既可以作为学术研究的基础平台,又可以作为科普的窗口,用户不需具备专家级的古典文献的阅读能力与丰富的历史知识便可以使用,大大便利了学术研究,特别是跨学科研究[5]。

目前,这种中国古籍的数据库的构建和应用仍存在三大问题亟待解决:①中国古籍要从字检索走向词检索。由于古汉语没有词语边界,要实现类似于英语的词检索功能,则必须进行词语的切分。例如,基于词检索“軍”时,只应该出现“軍”作为一个词的上下文,而不应该出现“將軍”“護軍”等结果。②从专名标引走向实体标注。不少经典古籍已经人工标引了人名、地名、书名等专名(专有名词),但是仅依靠横线和波浪线两种符号难以区分出不同类型的专名,更难以解决同名异指(如多个事物名称相同)和异名同指(如一个人物多个名称)问题。例如,检索“秦始皇”这个人物时,不仅要包含“秦始皇”这个字符串的上下文,还应该得到这个人物的所有上下文,包括“嬴政”“吕政”等。因而,全面梳理各种不同的人物、地点、时间等专名对应的实体信息,并在文本中标明每个实体的唯一代号,才能够满足后续细致的检索和统计需要。③从全文检索走向知识挖掘与可视化呈现。现有的检索平台大都只提供基于字的检索结果,而在人物、地点、时间等实体要素进行标注之后,可以借助数据挖掘技术,发掘出实体之间的关联关系,再通过可视化方法直观地展现出来。因此,需要探索能够实现基于词和实体

* 本文系江苏省社会科学基金项目“人工智能辅助青少年传统文化教育研究”(项目编号:20JYB004)、国家社会科学基金项目“中文抽象语义库的构建及自动分析研究”(项目编号:18BYY127)和国家社会科学基金重大项目“基于《汉学引得丛刊》的典籍知识库构建及人文计算研究”(项目编号:15ZDB127)研究成果之一。
作者简介:常博林,本科生;万晨,硕士研究生;李斌,副教授,博士,通讯作者,E-mail:libin.njnu@gmail.com;陈欣雨,本科生;冯敏萱,副教授,博士;王东波,教授,博士生导师。
收稿日期:2021-05-27 修回日期:2021-09-18 本文起止页码:134-142 本文责任编辑:易飞

的检索与知识挖掘的人文知识库构建方法。

本研究以文史价值极高的《资治通鉴·周秦汉纪》作为样本，构建数字人文知识库和检索系统。为了解决传统的基于字符串的全文检索存在的问题，将文本进行了分词与词性标注，从而实现基于词的全文检索。然后，进一步标注人物、地点的实体信息，并根据这些信息，借助可视化等技术，构建《资治通鉴》数字人文检索系统。在此基础之上，对人物、地点实体以及词汇进行计量与数据挖掘，给人文学者提供一个突破传统研究路径的更加高效的古籍信息加工框架与深度开发路径。

2 研究现状

《资治通鉴》以编年体方式记载了公元前 403 年至公元 959 年的历史，是一部史学与文学价值极高的典籍，其研究多集中于版本、点校、注疏和文学历史等方面。早在 1956 年，古籍出版社就出版了《资治通鉴》标点本[6]，1988 年董志翘对《资治通鉴》的标点提出疑误[7]。注疏方面分为 3 类：专题注释、节本注释和全书注释[8]。陈剩勇从政治功能和伦理功能角度评估了史学功能[9]，赵正阳从史学观角度概述了其史学价值和贡献[10]。

古籍的数字化研究工作与基于字符的全文检索已经成熟，有一批古籍全文检索数据库[1]。特别是 2014 年，中华书局推出了出版级的高质量《中华经典古籍库》[11]，收录了《资治通鉴》，功能有阅读全文、纪年换算、人名索引。专名也进行了标引，加上了专名线，例如，人物、地点、官职名、民族名等加下划线，书名加波浪线。

古文分词与词性标注也不断展开[12]。古汉语虽然以单字词为主，但是多字词仍占了相当的比例，大量的人名、职官、时间等均存在大量的多字词。分词之后，才可能实现词的检索。而名词、动词、人名、时间等细类区分的词类标注，对于古汉语的研究有重要意义。对于古籍检索来说，也可以更好地区分一个词的不同词类。由于建设成本高，目前仅有千万字级的语料库。主要有南京师范大学的先秦语料库[13]、中古汉语语料库[14]和台湾“中研院”的上古、中古、近代语料库[15]。

基于知识本体（ontology）方法的古籍内容结构化工作也已经展开。中华书局主持开发了“二十四史”本体，以人名索引、人名词典等资源，对 4 700 万字的二十四史中的人物、时间、地点等实体进行了自动提取和本体构建[16]。2007 年，北京大学数据分析研究中心团队与中华书局合作，设计开发了“资治通鉴知识服务系统”[17]。该系统通过对时间、地点、人物等专有名词进行标引，进而对人物进行相关性分析、时间分析等，是利用计算机分析技术对传统古籍进行知识挖掘的成功探索。2010 年，彭炜明、宋继华采用模式驱动的方式，构建了《资治通鉴》领域知识本体，并在此基础之上实现了本体的查询和可视化[18]。该项目更关注人物和事件的标注，但缺乏对地点的标注与分析。这两项《资治通鉴》知识库的开发，均使用了知识本体技术。不过，由于自动提取信息，导致实体的遗漏率较高，也没有解决好异名同指和同名异指问题。对语言学词汇信息标注（如分词、词类信息）和地理 GIS 信息等标注不足，有待更加全面的信息。

近年来，数字人文逐渐成为国内外人文研究的新方法。哈佛大学和复旦大学等合作开发了“中国历史地理信息系统 CHGIS”，提供了一个可以进行空间分析和时间统计的数字地图平台[19]。北京大学通过数字人文手段研究了唐代 300 年仕人的迁徙路线、宋到明几百年的儒家理学传承路线，开发了禅宗法传承可视化平台[20]。南京师范大学开发了《左传》[21]《史记·本纪》[22]两个包含词汇、人名、地名等实体与 GIS 信息的历史人文知识库，可以满足更为多样的检索与知识服务。

综上，古籍的电子化与字符级全文检索已经成熟，古文的分词与词性标注方法接近成熟，知识本体构建与分析逐步展开，基于词的全文检索成为未来主流的发展方向，时间、人物关系、GIS 等信息也越来越受到重视。《资治通鉴》等古籍专书知识库的建设亟需建立基于文本的实体标注，以实现更完整的信息整合与更多样、更深入的知识计量、挖掘与服务。

3 《资治通鉴·周秦汉纪》数字人文知识库的构建

《资治通鉴》的篇幅巨大，本文选取了最前面的周、秦、汉三代的数据进行建设，目的是为了先解决最早期的部分，并可以与记载内容相似的《史记》《左传》进行对比分析。考虑到基于字的全文检索或自动构建知识本体存在的问题，本研究尝试基于词语和实体的、地毯式的全文标注，以整合更多的信息，进行知识挖掘与可视化。实体标注目前仅限于人物和地点。表 1 给出了全文标注的 3 个层次，在原始文本的基础上，进行

词语的切分(用空格作为词界)、词性标注(名词、动词、标点等)和实体 ID(编号)的标注。这样每句话中的每个词都有了丰富的信息,通过标明人名、地名的 ID 号,解决同名异指和异名同指的问题。人名和地名对应的 ID 分别取自人物信息表和地名信息表,并与《左传》《史记・本纪》知识库中的实体 ID 保持相通。在标注时,沿用两者的人物实体表中的数据,新数据则分配新的 ID 进行信息填写和标注。下面分别介绍。

表 1 文本的多层次标注

标注层级	样例
原始文本	張耳、陳余至邯鄲
分词标注	張耳、陳余 至 邯鄲
词性标注	張耳[人名]、[标点] 陳余[人名] 至[动词] 邯鄲[地名]。[标点]
实体 ID 标注	張耳[人物 ID 3171]、陳余[人物 ID 2465] 至 邯鄲[地点 ID 981]

3.1 数据来源

《资治通鉴》的底本为繁体字,电子版全文 294 卷,总字数约 300 万字。本研究主要参照中华书局 1956 年本[23]进行校勘。目前,完成了周、秦、汉 3 个朝代共计 68 卷(60 万字)的文本校勘与标注工作。

3.2 分词与词性标注

古文分词和词性标注工作,耗时耗力。本文采用了机器自动标注,然后辅以人工校正的方式,进度大为加快。首先,采用了陈小荷等制定的分词与词类标记集[13],使用南京师范大学古汉语词性标注系统[24]进行了自动分词与词性标注,该系统的整体正确率在 85% 以上,然后进行了全面的人工校对,形成高质量的标注文本。

3.3 实体信息标注

3.3.1 人物信息

《资治通鉴》中人物的名号往往有多个,并且不同人物的同名现象也相当普遍,需根据各种注疏文献和相关资料进行辨析。为了辨识清楚每个人物,本文给每个人物实体分配一个唯一的 ID 号(即编号)。如果这个人物在《左传》和《史记》出现过,则沿用这两部书的人物 ID。对于新的人物,则设立新的 ID。人物信息还包括人物的各种名称、性别和国别。由于一个人物在古书中名称可能较多,为了便于后续的检索和可视化显示,我们还设置了后世使用较多的"人物主名"作为人物的正名。"人物主名"并不一定来自"人物名",而可能是后世采用的较为完整的名称。如表 2 所示,"叔孙州仇"的人物 ID 为 131,有 4 个名字,性别为男,国别为鲁。

表 2 人物实体示例

人物 ID	人物主名	人物名	性别	国别
131	叔孙州仇	叔孙 武叔懿子 州仇 子叔孙	男	鲁

3.3.2 地点信息

与人物信息标注相似,地点也沿用了《左传》和《史记》中的信息,对于《资治通鉴》中新出现的地名,则予以新的 ID,并填写地理实体的信息,包括地名的类别(国家、诸侯国、河流、山川等)、今天的所在地、考据的文献出处,然后根据今天所在地查出百度地图的地理 GIS 坐标。主要参考《中国历史地图集》[25]、中国历史地理数据库 CHGIS[19]等资料。表 3 给出了诸侯国"邾"的基本信息。

表 3 地理实体示例

地名 ID	地名	类别	今天的所在地	考据的文献出处	百度 GIS 坐标
2	邾	诸侯国名	山东省邹城市东南	杨伯峻《春秋左传注》	117.008 519, 35.413 84

3.3.3 时间信息

根据《先秦诸子系年》等资料[26],将每一个篇目的年号,对应到公元纪年上。例如,"卷第一・周紀一・二十一年"对应于"公元前 381 年"。

3.4 数据库架构

基于《资治通鉴》电子化全文、分词和词性标注以及实体信息标注,构建出《资治通鉴・周秦汉纪》数据库。主要包括人物实体、地点实体表、文本表、标注文本表、人物同现表、人地同现表共计 6 张数据表,具体字段与结构如图 1 所示。根据人物实体表和地点实体表中的 ID,将正文中的每个人物和地名都标注了其 ID 信息。然后,同一个句子中,不同的人物会一起同现,人物和地点也会同时出现。我们根据这两种同现信息,在标注好的"标注文本表"上,提取出"人物同现表"和"人地同现表"。

4 基于词和实体的全文检索

4.1 基于词和实体的检索框架

为了让平台服务社会,本研究使用 Web 开发技术,构建了《资治通鉴》在线检索系统,测试版网址为 www.dhbase.com/zztj。系统的功能结构如图 2 所示,除了基于词的全文检索功能外,还基于底层的结构化的数字人文知识库,提供了人物、地点、词性等多种查询方式。

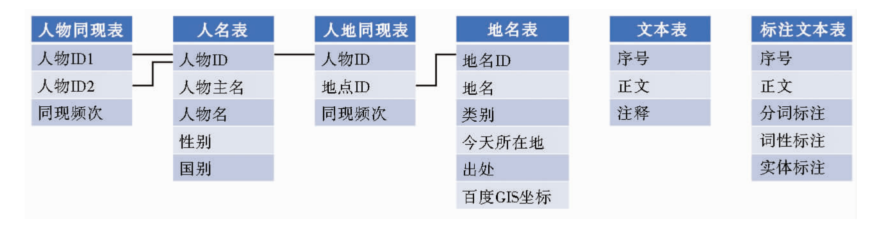


图 1　数据库的结构

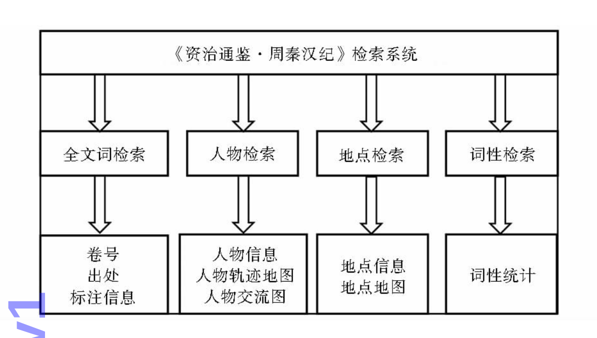


图 2　《资治通鉴》检索系统结构

4.2　全文实体检索

与传统的基于字符串匹配的全文检索方式不同，全文实体检索是建立在具有实体标注信息的文本之上的一种更加精确的、基于词和实体的检索方式。这种方式避免了字符检索生硬匹配造成检索冗余、缺漏与错配问题。例如图 3 给出了“軍”基于词的结果，就不会再出现包含在“將軍”“北軍”“護軍”等词中的情况，从 2 098 个结果，减少到 1 872 个。



图 3　基于词的检索结果(“軍”)

4.3　人物检索

人物检索功能主要提供了人物的基本信息的查询，主要包括人物的主名、别名、性别、国别信息，并且能与《左传》和《史记·本纪》知识库进行联动，展示该人物在《左传》和《史记·本纪》中的出现情况。例如在检索“汉武帝”时，系统根据人物 ID 进行全文检索，可以得到“武帝”“刘彻”等结果，而不仅限于包含“汉武帝”的段落，如图 4 所示：



图 4　人物检索示例

此外，人物地图检索功能基于人物与地点的同现数据，借助百度地图实现人物可能去过的地点展示，并基于人物与人物的同现数据，借助 ECharts[27] 技术实现人物交际圈的图示。图 5 给出了汉武帝的人物主名、别称、性别、国别以及在《史记·本纪》和《左传》中出现的情况。



图 5　人物同现地的地图示例(“汉武帝”)

图 6 用人物同现数据近似展现了汉武帝的交际情况，图的中心节点表示汉武帝，周围节点表示与汉武帝

在一个句子中同现的人物。在这些同现人物节点中，节点越大表示该人物与汉武帝同现的次数越多，人物关系就可能越密切。

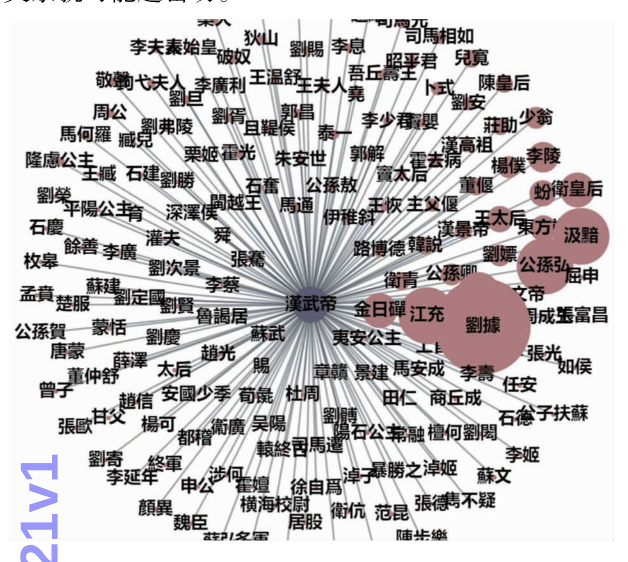


图 6 “汉武帝”的人物关系

4.4 地点检索

地点功能提供了地点实体的基本信息的查询，主要包括地点的地名、类别、今天所在地，并且能与《左传》和《史记・本纪》知识库进行联动，展示该地点在《左传》和《史记・本纪》中的出现情况。此外，地点检索功能基于地点实体表中的坐标信息，借助百度地图实现地点在现代地图上地理位置的展示。例如，在检索“长安”时，可以看到长安的地名、类别、今天所在地等信息，并能看到长安在现代地图上的具体位置。

4.5 词性检索

除了人物和地点的实体检索之外，普通词语也都进行了分词和词性标注，可以在词性信息的支撑下，提供基于词性的检索和统计。词性统计功能提供了词性的相关信息，主要包括该词性下的所有词条，和借助 ECharts 绘制的该词性不同词条的词频条形图。图 7 展示了检索“v（动词）”得到的词频分布图，可以方便古汉语研究者查看词性信息。

5 计量分析与知识挖掘

根据《资治通鉴・周秦汉纪》数字人文知识库和检索系统，可以进行更加深入的计量分析工作，获得传统定性分析难以得到的数据结果。在《资治通鉴・周秦汉纪》所包含的实体中，人物实体有 4 588 个，地点实体有 1 451 个。人物统计结果显示，平均每个人物有 1.95 个名称，有超一半的人物不止 1 个名称，50% 的人物有 2－5 个名称，3% 的人物有 6 个及以上的名称。名称最多的人物是汉高祖刘邦。因此，给每个人物使用唯一的 ID 编号来标定是十分有必要的。下文将针对词语、人物与地点实体展开大数据计量统计与可视化分析。

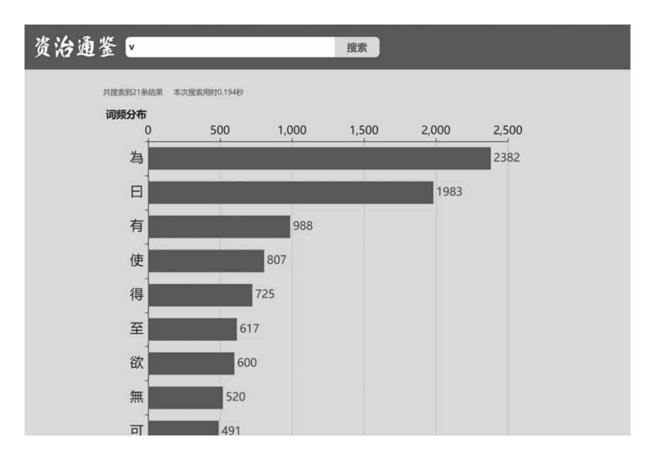


图 7 词性检索“v（动词）”示例

5.1 词频统计

不同于传统的文史分析，数字人文视域下可以利用能够揭示或表达文本核心内容的关键词或主题词出现的频次高低来确定其主题和热点。基于已标注的分词文本，可以得到词频统计结果。统计结果显示，单字词共有 2 610 个，多字词共有 7 970 个。其中“之”一词出现的频率最高，为 5 038 次。不同于基于字的传统研究，本研究在基于词的语料上，可以统计出多字词的分布情况。表 4 列出了词频排名前 10 的多字词，这些词多为实词，且基本都和争夺天下相关，例如“天下”“諸侯”“陛下”“將軍”等，在一定程度上反映了《资治通鉴》主要是围绕争夺天下与帝位记叙的。其中“天下”一词出现的频率最高，为 569 次。当然，还可以利用词语的信息，从历时的角度来观察汉语词汇演化的过程，可以研究词汇的动态发展状况。

表 4 多字词的词频（排名前 10）

多字词	词频
天下	569
諸侯	275
匈奴	267
太子	245
天子	237
於是	209
陛下	209
將軍	169
漢王	162
丞相	149

5.2 实体关系的挖掘与 3 部史书横向对比

5.2.1 最“广交”人物

在传统的人物研究中，人物历史地位的高低往往

需要根据其在历史事件中所扮演角色的轻重来衡量。这种评价方式虽然可行,但却缺乏统一且客观的标准。而通过定量统计人物同现次数的方式,可以近似地估计人物之间的交往关系,进而估计人物的历史地位。同现人物越多,交际也就越广,所具有的地位也就可能更高。横向对比《左传》和《史记·本纪》的数据,能够明显地看出3本史书的异同。为了更好地和《史记》对比,我们将《资治通鉴》的数据截止到汉武帝时期。表5列出了《资治通鉴·周秦汉纪》最“广交”的十大人物,其中前3名为汉高祖、汉武帝和项羽。3本书相对照,可以看出《史记·本纪》与《资治通鉴·周秦汉纪》更偏重对秦汉时期的记载。

表5　最广交人物(前10位)

资治通鉴·周秦汉纪			左传		史记·本纪	
人物ID	人物	人物数	人物	人物数	人物	人物数
2 621	汉高祖	168	晋文公	99	汉高祖	109
2 625	汉武帝	164	晋悼公	85	项羽	71
3 079	项羽	82	范宣子	71	汉惠帝	36
2 797	吕太后	60	晋景公	70	项梁	32
2 618	汉文帝	57	楚庄王	65	吕太后	31
2 623	陈涉	52	齐桓公	62	帝舜	29
2 624	秦昭王	52	郑文公	61	黥布	27
2 873	汉景帝	47	晋厉公	57	韩信	26
2 868	韩信	47	羊舌肸	56	陈平	25
3 171	秦始皇	42	楚共王	55	刘肥	25

5.2.2　人物游历距离

《资治通鉴·周秦汉纪》记录了大量的时间、人物和地点信息。借助人物与地点实体的同现信息可以近似地估计人物可能的游历地点。结合地点的坐标信息可以计算出两点A(纬度φ_1,经度λ_1)和B(纬度φ_2,经度λ_2)之间的球面距离[28],如公式(1)所示。加上时间的先后顺序,将各个距离累加起来,可以估计人物的游历距离。

$$\text{Distance}(A,B) = 111.999 \times \sqrt{(\varphi_1-\varphi_2)^2+(\lambda_1-\lambda_2)^2\cos^2\left(\frac{\varphi_1+\varphi_2}{2}\right)} \quad \text{公式(1)}$$

如表6所示,在《资治通鉴·周秦汉纪》游历距离最多的10位人物中,4位为君王,3位为军事家,2位为开国元勋,1位为外交家。其中游历距离最多的汉高祖,距离有14万千米之多,可见汉高祖征战开国的一生。此外,通过联动《左传》和《史记·本纪》数据库,也可以比较不同史书中人物游历的差异,进而挖掘两本史书在内容和风格上的不同倾向性。借助这种方法,虽然不能进行精确的计算,但能大致地估计出人物游历的轨迹与行程,辅助分析人物的生平、出行距离等问题。

表6　人地同现数及距离(排名前10)

资治通鉴·周秦汉纪			左传			史记·本纪		
人物	地点数	直线距离/千米	人物	地点数	直线距离/千米	人物	地点数	直线距离/千米
汉高祖	933	146 288	周武王	48	19 300	汉高祖	253	33 393
项羽	533	79 493	崔杼	33	16 018	项羽	226	20 875
汉武帝	261	70 192	晋文公	47	15 964	舜帝	30	15 151
韩信	324	34 739	楚庄王	36	14 907	黄帝	36	12 813
张骞	38	33 733	范宣子	35	14 498	韩信	61	12 431
彭越	175	31 079	郑文公	39	14 362	章邯	55	12 153
吕太后	72	27 045	秦穆公	33	13 860	秦始皇	45	11 996
刘安	48	23 297	知武子	36	13 828	黥布	36	10 164
张耳	145	22 233	晋景公	38	12 594	彭越	53	7 255
陈余	131	22 186	周文王	33	11 996	项梁	56	7 170

5.3　实体历时统计分析

5.3.1　多事之秋——实体历时分布

从文本中实体所出现的频次密度的角度来分析可以更好地呈现不同时间段之间的差异性。将《资治通鉴·周秦汉纪》所记载实体分别对应到公元纪年法,可以得到其所对应的公元前403年至公元前87年间的实体曲线。如图8所示,蓝色曲线表示相应时间的人物数量,橙色曲线表示相应时间的地点数量。可以发现,在整个时间区域内,人物略多于地点,且存在时间差异性。人物和地点均在公元前207年前后达到峰值,反映了历史上具有重大决战性的巨鹿之战;人物和地点曲线在公元前154年前后同时上升,反映了历史上西汉规模最大的一次诸侯王国叛乱——七国之乱。通过在时空角度进行分析的方法,可以快速地定位发生重大事件的历史时代。

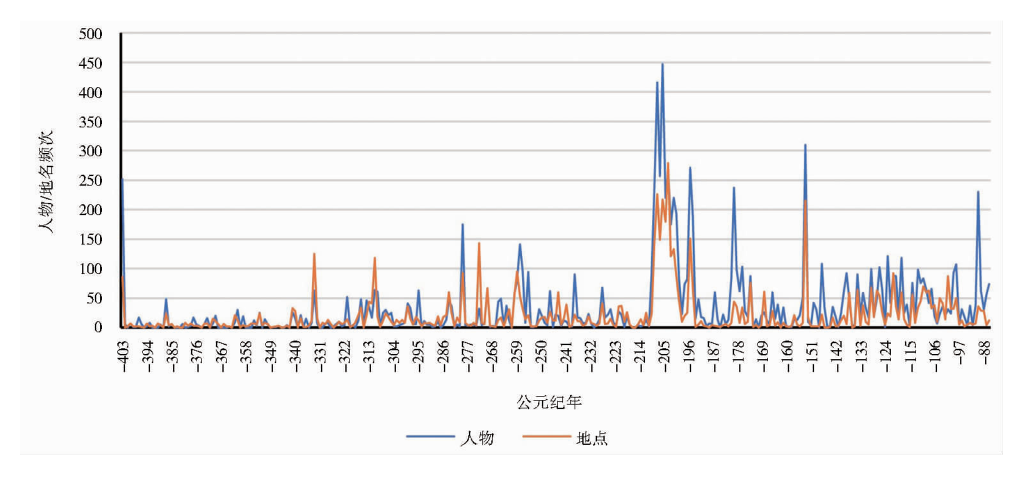


图 8 人物与地点历时分布

5.3.2 风云人物与风云之地——特异性实体挖掘

《资治通鉴》是一部编年体通史,时间信息丰富。利用时间信息,可以挖掘出每个时代的文本中的特异成分。例如,可以利用特异性指标 TF-IDF,来得到在不同时代举足轻重的风云人物与和风云之地。

TF-IDF (Term Frequency - Inverse Documentation Frequency)算法是由 G. Salton. 提出来的用于信息检索的算法[29]。TF-IDF 主要基于一个思想,即词区分特定文本内容的能力随着其在该文本中出现的频率的升高而提升(TF),随着所出现文本的范围的扩大而下降(IDF)。因此,TF-IDF 越高,表明该词汇的文本区分度或者说特异性越强;TF-IDF 越弱,表明该词汇的文本区分度或者说特异性越弱。

通过 TF-IDF 算法分析,可以得到一批具有时代特色的人物实体,进而借助上一节的人物历时分析方法,可以将每个时代的重要人物绘制在时间轴上。利用 ECharts 的流体图可视化工具,绘制出图 9,可以看出一个个重要的历史人物呈现出"你方唱罢我登场"的态势。吴起自公元前 412 年被鲁元公起用而走入历史舞台,至公元前 381 年被贵族射杀而淡出历史记载;秦始皇嬴政自公元前 259 年出生被历史关注,至公元前 210 年逝世走出历史视野;汉文帝自公元前 178 年文景之治的开始达到顶峰,汉景帝至公元前 141 年逝世结束文景之治。通过风云人物实体的挖掘,能够更直观地看到历史的演化。

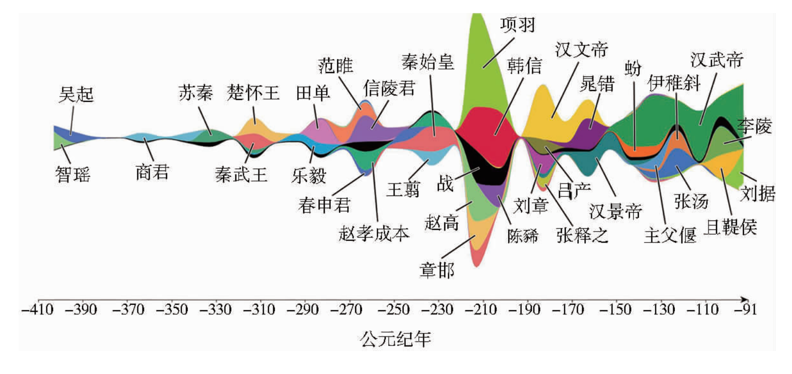


图 9 风云人物历时演变

同样地,以相同的方式可以得到不同时代最重要的地点流变,如图 10 所示。可以发现,不同的地点随着历史的演变而此起彼伏地成为一个个时代的焦点,或是诸侯国都,或是兵家必争之地,都成为时代更迭的印记。

6 结语

在古籍数字化与全文检索已普遍应用的今天,如何综合运用数字人文的新技术和手段,充分利用我国种类内容丰富的历史文献资源,在全文检索的基础上进行可视化呈现和大数据分析,是当今文学、历史和图书情报等领域的重要课题。本研究在数字人文的研究范式下,针对基于字的全文检索存在的词语边界和实体概念不明问题,以及本体知识库与原文脱节问题,提出使用全文词语标注的解决方案,尝试建设了《资治通鉴·周秦汉纪》数字人文知识库,对文本进行了词语切

分、词性标注和实体信息的全文标注。其次,开发了基于词和实体的全文检索系统,包括人物检索、地点检索、词性检索等,并借助百度地图和 ECharts,可视化地展现了相关的人物游历、地理信息和人物关系。然后进行了计量分析与知识挖掘,穷尽统计了《资治通鉴·周秦汉纪》中的人物数量。对于实体进行了多角度的关联分析与挖掘,例如人物交际、人物游历地图、多事之秋、风云人物地点等。还通过与《左传》《史记·本纪》的比较,统计出 3 本书记述的人物差异。

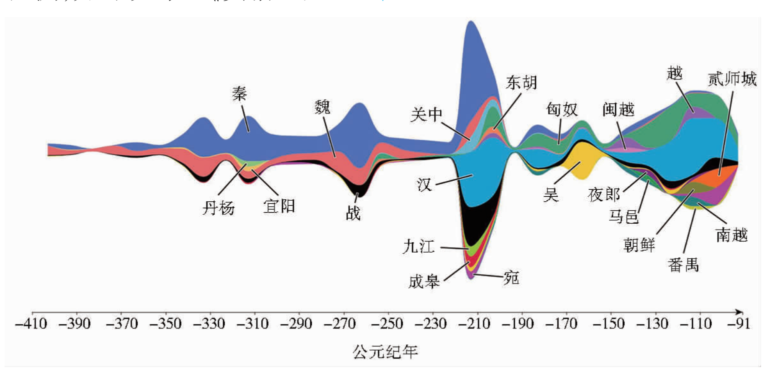


图 10　风云之地历时演变

在未来的工作中,我们将继续做好以下研究工作:①扩大数据规模,将整部《资治通鉴》标注完整,并反复检查校正。②继续完善实体信息标注,结合学界最新的考据成果,不断修订数据库的内容。同时,还需提高系统开放度,增加纠错与审核机制,吸引更多的专家学者参与到项目工作中来。实体标注的对象也可以扩展至官职、年号、器物等更多类型。③探索更多的知识统计与知识挖掘方法。实体之间的同现当前还只是处于近似估计的阶段,未来也可以优化计算方法,提高计算的精确度。此外,还可以考虑对人物关系和人地关系进行更细致的分类。例如人物关系增加朋友、亲属、上级下属等。④改进检索与数据可视化方式。如将当前不同的检索方式有机地进行整合,提高检索效率,提升当前可视化的展示效果。⑤此外,还可以同图书馆、博物馆数据库联通,将《资治通鉴》的文本信息与其他历史文献和馆藏品进行链接,将诸多历史要素融于一体进行呈现。

参考文献:

[1] 季培培.常见 10 种古籍全文数据库的比较研究[J].图书馆学研究,2020(20):71－80.

[2] 刘炜,叶鹰.数字人文的技术体系与理论结构探讨[J].中国图书馆学报,2017,43(5):32－41.

[3] The Open University. Hestia[EB/OL].[2021－05－21]. https://hestia.open.ac.uk/.

[4] 中国历代人物传记数据库管理委员会.中国历代人物传记数据库项目(China Biographical Database,CBDB)[EB/OL].[2021－05－21]. https://projects.iq.harvard.edu/chinesecbdb.

[5] 欧阳剑.面向数字人文研究的大规模古籍文本可视化分析与挖掘[J].中国图书馆学报,2016,42(2):66－80.

[6] 宋衍申.试探建国以来的《资治通鉴》研究[J].东北师大学报,1983(5):88－93.

[7] 董志翘.《资治通鉴》标点疑误[J].古汉语研究,1988(01):83－87, 36.

[8] 林嵩.南宋《通鉴》注考论[J].古代文明,2007(1):74－81, 113.

[9] 陈剩勇.资治通鉴:中国传统史学功能分析[J].史学理论研究,1995(4):74－80,146.

[10] 赵正阳.司马光《资治通鉴》的概述及其史学价值[J].北方文学,2019(9):41－42.

[11] 中华书局.中华经典古籍库[EB/OL].[2021－05－21]. http://publish.ancientbooks.cn/docShuju/platformSublibIndex.jspx?libId=6.

[12] 邓三鸿,胡昊天,王昊,等.古文自动处理研究现状与新时代发展趋势展望[J].科技情报研究,2021,3(1):1－20.

[13] 陈小荷,冯敏萱,徐润华,等.先秦文献信息处理[M].北京:世界图书出版公司,2013.

[14] 王晓玉.中古汉语语料库的设计与实现[J].辞书研究,2017(3):17－26.

[15] 台湾"中研院"古汉语标注语料库[EB/OL].[2021－05－21]. http://lingcorpus.iis.sinica.edu.tw/cgi-bin/kiwi/akiwi/kiwi.sh.

[16] 董慧,徐雷,王菲,等.语义分析系统研究(Ⅲ)——中华史籍语义分析系统实现[J].情报学报,2014,33(2):204－214.

[17] 孙显斌.基于本体的古籍分析系统开发实践——以"资治通鉴分析系统"为例[C]//科学数据管理、仓储和应用实践研讨会论文集, 2019.

[18] 彭炜明,宋继华.《资治通鉴》历史领域本体构建及其应用研究[J].中文信息学报,2010,24(2):33－38.

[19] 中国历史地理信息系统 CHGIS[EB/OL].[2021－05－21]. https://sites.fas.harvard.edu/~chgis/.

[20] 严承希,王军.数字人文视角:基于符号分析法的宋代政治网络可视化研究[J].中国图书馆学报,2018,44(5):87－103.

[21] 李斌,王璐,陈小荷,等. 数字人文视域下的古文献文本标注与可视化研究——以《左传》知识库为例[J]. 大学图书馆学报,2020,38(5):72-80,90.
[22] BIN L, YAXIN L, QIAN Y, et al. From history book to digital humanities database: the basic annals of the Shiji[J]. Journal of Chinese history, 2020, 4(2): 528-536.
[23] 司马光. 资治通鉴[M]. 北京:中华书局,1956.
[24] 石民,李斌,陈小荷. 基于 CRF 的先秦汉语分词标注一体化研究[J]. 中文信息学报,2010,24(2):39-46.
[25] 谭其骧. 中国历史地图集[R]. 北京:中国地图出版社,1982.
[26] 钱穆. 先秦诸子系年[M]. 北京:商务出版社,2015.
[27] Apache Software Foundation. ECharts[EB/OL]. [2021-05-21]. https://echarts.apache.org/zh/index.html.
[28] 韩忠民. 知经纬度计算两点精确距离[J]. 科技传播,2011(11):196,174.
[29] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information processing and management, 1988,24(5): 513-523.

作者贡献说明:
常博林:软件设计编写,数据统计,初稿撰写;
万晨:数据标注和校对,初稿撰写;
李斌:总体思路设计,数据核对,论文修改;
陈欣雨:数据标注和校对;
冯敏萱:数据组织校对,论文修改;
王东波:理论论述,论文结构调整,修改论文。

The Construction and Application for Digital Humanities Knowledge Base of Ancient Books Based on Word and Entity Annotation: A Case Study on *Zhou Qin Han Annals of Zizhitongjian*

Chang Bolin[1] Wan Chen[2] Li Bin[1] Chen Xinyu[1] Feng Minxuan[1] Wang Dongbo[3]
[1] School of Chinese Language and Literature, Nanjing Normal University, Nanjing 210097
[2] Department of Chinese language and literature, Fudan University, Shanghai 200433
[3] College of Information Management, Nanjing Agricultural University, Nanjing 210095

Abstract: [**Purpose/significance**] To explore a humanistic knowledge base construction method based on word and entity retrieval and knowledge mining. [**Method/process**] This paper constructed the *Zhou Qin Han Annals of the Zizhitongjian*, achieved the automatic segmentation and part-of-speech tagging of the 68-volume 600,000-character text, manually annotated entity information such as persons, locations, GIS and time in the text, and designed the system of full-text retrieval and map visualization based on words and entities. This paper used co-occurrence information to get the relationship and travel information of the characters. By TF-IDF and time series analysis, the key periods, people and locations in history were automatically extracted and illustrated. [**Result/conclusion**] Depth information labeling based on words and entities is a good solution to the problems of word boundaries, same name with different person and different name with same person, and it can solid the basis for multi-studies on the knowledge mining and knowledge service of ancient books.

Keywords: *Zizhitongjian* digital humanities knowledge mining ancient book retrieval ancient Chinese language processing